Universität Bonn
Abteilung für Romanische Philologie
Sommersemester 2024
Sprachanalyse im technologischen Wandel:
Eine Einführung in die Computerlinguistik (Frz./Ital./Span.)
Iris Ferrazzo M.A.

Zwischen den Zeilen lesen:
Können domänenspezifische Embeddings Hinweise auf politische Überzeugungen und Vorurteile geben?

22.08.2024

vorgelegt von

Simon Köhl
Informatik, 6. Fachsemester
50004834
Endenicher Straße 279
53121 Bonn
s03skoeh@uni-bonn.de

Inhalt

1. Einleitung

Embeddings gewannen in den letzten Jahren immer mehr an Bedeutung (Gavagai, 2020), und das zurecht: Durch sie eröffnen sich damals wie heute zahlreiche Türen zur Analyse, Verarbeitung und zum Verständnis von natürlicher Sprache. Sie erlauben es uns, Worte als Punkte in einem hochdimensionalen Vektorraum darzustellen, in welchem mathematische Methoden zur Verfügung stehen. So können semantische und syntaktische Beziehungen zwischen Wortfeldern betrachtet (Mikolov et al., 2013 (1)), große sprachliche Veränderungen über Zeit (Hamilton et al., 2018) und Domain (Azarbonyad et al., 2017) erkannt und sogar automatische Übersetzung gemeistert werden (Mikolov et al., 2013 (2); Joulin et al., 2018).

In der Analyse von politischen Reden versprechen Embeddings ebenfalls beachtliche Vorteile. Azarbonyad et al. (2017) erkannten mithilfe derer unter anderem semantische Shifts in politischen Texten bezüglich verschiedener Schlüsselworte.

In diesem Report sollen ebenfalls Unterschiede in der Struktur der Embeddings bezüglich politischer Texte untersucht werden. Das Ganze soll am Fallbeispiel von Reden im Bundestag von 2018 bis 2022 der Parteien Alternative für Deutschland (kurz: AfD) und Bündnis 90/Die Grünen (auch: Die Grünen) geschehen. Dabei wird gezeigt, wie bestimmte politische Ausrichtungen trotz rhetorischer Maskierung analytisch nachgewiesen werden können. Dafür soll zuerst knapp auf verwandte Arbeiten eingegangen, daraufhin die Motivation dieses Reports und einige Terminologien erörtert werden. Anschließend wird die Methodologie im Detail beschrieben. Es wurden Embeddings beider Parteien erstellt, auf welchen einige quantitative Analysen ausgeführt wurden. Danach werden die Ergebnisse erläutert, welche im darauffolgenden Kapitel bezüglich ihrer Verlässlichkeit auf Basis der verwandten Methoden eingeordnet werden sollen. Schließlich wird als Fazit eine Zusammenfassung des Reports geliefert und ein Ausblick auf mögliche Verbesserungen gegeben.

2. Verwandte Arbeiten

Embeddings als Anhaltspunkte für Analysezwecke sind, trotz Ihrer mathematischen Mächtigkeit, in der Forschung vergleichsweise unterrepräsentiert. Dieses Kapitel wird sich kurz mit zwei verwandten Arbeiten zu diesem Report

beschäftigen, welche alle Embeddings nutzen und auf denen Teile der Methodologie dieses Reports aufbauen.

2.1 Erkennung semantischer Shifts durch diachrone Wort-Embeddings

Die Analyse semantischer Shifts sucht nach Worten, deren Kontext sich über Zeit und Domain verändert. Hamilton et al. (2018) untersuchten anhand mehrerer Korpora der englischen Sprache zu verschiedenen Zeitepochen eben genau diese semantischen Shifts, indem zuerst Embeddings für jeden Korpus erstellt wurden. Diese wurden dann mithilfe des orthogonalen Procrustes so ausgerichtet, dass die paarweise Distanz der Punkte in den Embeddings minimal wird (diese Methode wird auch *Alignment* genannt). Die aneinander ausgerichteten Embeddings konnten daraufhin miteinander verglichen werden, indem die Verschiebung eines Wortes über verschiedene Zeitepochen berechnet wurde. Zudem wurden andere Metriken wie zum Beispiel die Anzahl übereinstimmender kontextueller Nachbarworte verwendet, um schließlich automatisch die Worte mit dem größten semantischen Shift zu entdecken. Das Forschungsteam evaluierte ihre Methode, indem sie bereits bekannte Shifts mit den automatisch entdeckten verglichen. So entdeckte ihre Vorgehensweise zum Beispiel den semantischen Shift des englischen Wortes „*gay*“, früher eher mit Freude und Glück konnotiert, zu einer Region entsprechend dem Kontext von Homosexualität. Die Autoren postulieren darauf aufbauend statistische Gesetze für semantischen Wandel, welche für diesen Report allerdings von niederer Bedeutung sind.

2.2 Erkennung semantischer Shifts in politischen Texten

Azarbonyad et al. (2017) bauen auf im vorangegangenen Unterkapitel genannten Arbeit auf, um semantische Shifts nicht nur über der Zeitdimension, sondern auch zwischen verschiedenen politischen Überzeugungen zu entdecken und analysieren. Sie stoßen hierbei auf das Problem, dass es sich nicht explizit um eine Sprache, sondern um ein gesondertes Vokabular pro politische Überzeugung handelt. Die Dimensionalität und das Vokabular der Embedding-Räume sind also unterschiedlich, das *Alignment* erschwert sich. Um dieses Problem zu lösen, werden Stopp- und besonders häufige Wörter als Fixpunkte ausgewählt, deren Kontext sich idealerweise nicht ändert. Für diese wird mittels Gradientenabstiegsverfahren eine Transformationsmatrix gelernt, welche die Embedding-Räume aneinander ausrichtet. Schließlich wird eine Definition für die Stabilität eines Wortes

eingeführt, auf Basis welcher die Worte erkannt werden können, deren semantische Bedeutung sich am meisten unterscheidet.

3. Motivation

Der politische Diskurs in Deutschland wird zunehmend komplizierter. Immer mehr finden Falschwahrheiten und „alternative Fakten“ Einzug in das gemeinschaftliche Denken und spalten die Meinungen der Bevölkerung (Informationsschreiben der Bundesregierung zur Erkennung von Falschmeldungen, 2024). Diese Meinungsspaltung und -unterwanderung hat unter anderem auch zur Konsequenz, dass ein Wort abhängig vom Rezipienten unterschiedliche Assoziationen auslöst, was zusätzlich das gegenseitige Verständnis und das aktive Zuhören erschwert. Spricht das Gegenüber von „Feminismus“ kann für den populistisch-konservativen Zuhörer die Konversation direkt eher negativer und linkspolitischer Natur sein. Somit ist es von großer Bedeutung, den Kontext verschiedener Schlüsselworte im politischen Diskurs einzuordnen und vor allem jene Worte zu erkennen, deren semantische Bedeutung sich in verschiedenen Meinungsbildern am meisten unterscheidet. Dieses Kapitel gibt einen kurzen Überblick über die Motivationen der verwendeten Konzepte und eine knappe inhaltliche Einführung in dieselben.

3.1 Über die Struktur von Embeddings

Embeddings bieten sich per definitionem für die Analyse semantischer Bedeutung an: Sie berechnen eine Vektor-Darstellung von Worten/Tokens, in der ähnliche Worte nahe beieinander liegen. Bekannte Methoden zur Berechnung von Embeddings sind Word2Vec (Mikolov et al. (1), 2013) in seinen Varianten der Continuous-Bag-of-Words- (CBOW) oder Skip-Gram-Modelle, GloVE (Pennington et al., 2014), welches globale anstelle von lokalen Vektorrepräsentationen berechnet, und BERT (Devlin et al., 2019). Von besonderem Interesse für diesen Report sind vor allem die Word2Vec-Modelle, spezifisch die Skip-Gram Variante, welche besonders gut semantische Beziehungen erfassen können (Mikolov et al. (1), 2013; Mikolov et al. (3), 2013).

Embeddings liefern also kontextuelle Informationen zu einzelnen Worten und erzeugen somit eine Punktwolke, die die semantischen Beziehungen aller verwendeten Worte und Phrasen repräsentiert. Um die Räume miteinander automatisch zu vergleichen, müssen sie allerdings zuerst aneinander ausgerichtet werden, ein Problem, welches das *Alignment* lösen soll.

3.2 Alignment

Zum Ausrichten zweier Punktwolken zueinander stehen diverse Methoden zur Verfügung (Biswas et al., 2020; Kalinowski und An, 2020). Die wohl bekannteste ist der Procrustes Algorithmus, welcher eine Matrix berechnet, die eine Punktwolke so rotiert, skaliert und schert, dass sie einer anderen möglichst genau entspricht. Die handlichere Variante hiervon ist der orthogonale Procrustes Algorithmus, mit dem eine orthogonale Matrix berechnet werden kann, welche die Transformation unter bestimmten Voraussetzungen ausführt. Weiterhin existieren Alignment-Methoden, welche sich Werkzeugen aus dem mathematischen Bereich des optimalen Transportes bedienen. Kalinowski und An (2020) bieten einen Überblick über gängige Methoden.

4. Methodologie

Die folgenden Abschnitte behandeln die Konstruktion der Korpora und wie aus denselben die Embeddings erzeugt wurden, mit denen schließlich die im letzten Abschnitt beschriebenen Analysen betrieben wurden. Der Code ist in Form von Jupyter Notebooks als .zip Datei der Abgabe beigefügt.

4.1 Korpus

Die Embeddings wurden aus zwei Korpora gelernt, jeweils bestehend aus politischen Texten der AfD- und Grünen-Fraktion des Bundestages. Die politischen Texte stammen aus dem Zeitraum von 2018 bis einschließlich 2022 und wurden von OpenDiscourse bereitgestellt (Limebit GmbH, 2021). Da sich hier zum Zeitpunkt des Verfassens dieses Reports nur maximal 200 Reden herunterladen ließen, wurden für jede Partei CSV-Dateien erstellt und anschließend in eine große Textdatei zusammengefügt. Reden mit einer Länge von weniger als 10 Zeichen wurden, direkt im Vorhinein, ausgefiltert. Daraus ergaben sich 5914 Reden im Korpus der AfD und 5500 Reden im Korpus der Partei Bündnis90/Die Grünen.

4.2 Preprocessing

Um eine möglichst gute Downstream-Performance zu erhalten, müssen die Korpora bereinigt werden. Dazu wurden zuerst Interpunktionen entfernt und die Reden tokenisiert, sodass ein Wort ein Token darstellt. Anschließend wurden bekannte Stoppwörter entfernt und die Tokens lemmatisiert, um eine Übersättigung von Nachbarschaften mit syntaktisch und grammatikalisch ähnlichen Worten zu

vermeiden. Alle Tokens wurden daraufhin in Kleinschreibung überführt und Zahlen wurden entfernt. Danach wurde ein Bigram-Phrasing angewandt, um beispielsweise Tokens wie „New“ und „York“ zu „New_York“ zu bedeutungserhaltenden Phrasen zusammenzuführen. Dieser Vorgang beruht auf statistischen Regeln und ist damit nicht fehlerfrei. Schließlich wurden gängige Floskeln, die anwendungsspezifisch eine übermäßige Frequenz hatten, ausgefiltert. Beispielsweise beginnt fast jede Rede mit „Sehr geehrte Damen und Herren, …“. Diese Begrüßungsfloskel trägt aufgrund ihrer universalen Anwendung kontextuell nahezu keine Bedeutung, weshalb sie hier vernachlässigt wurde. Für jeden Korpus wurden anschließend einige linguistische Statistiken erhoben, welche in Abb. 1 und Tabelle 1 aufgeführt sind.

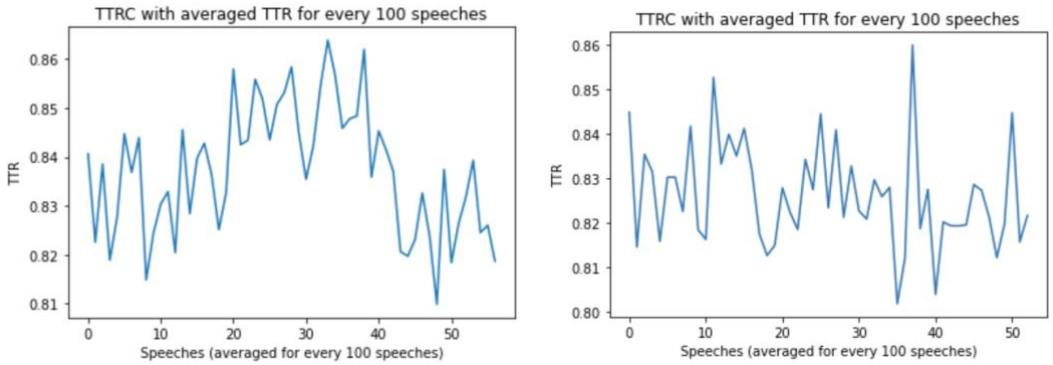

Abbildung 1: TTRCs der einzelnen Fraktionen, gemittelt über Intervalle von jeweils 100 Reden für bessere Lesbarkeit. Links: TTRC der AfD-Fraktion. Rechts: TTRC der Grünen-Fraktion

Tabelle 1: Statistische Kenngrößen der einzelnen Korpora

Partei	Types insgesamt	Tokens insgesamt	TTR über gesamten Korpus (gerundet auf 7 Nachkommastellen, nicht normalisiert)	STTR, visuell aus TTRC bestimmt
AfD	1062457	81139	0.0763692	ca. 0.843
Bündnis 90/Die Grünen	932016	63748	0.0683980	ca. 0.832

Wichtig ist hier zu nennen, dass das standardisierte TTR nur visuell aus den TTRCs bestimmt wurde, und somit keine hohe Qualität der Werte zu erwarten ist.

In der weiteren Analyse werden diese Werte allerdings auch nicht weiter betrachtet. Allgemein lässt sich dennoch sagen, dass das TTR bei beiden Parteien ähnlich ist.

4.3 FastText

Als Embedding-Modell wurde FastText (Bojanowski et al., 2017) verwendet. FastText ist eine Variante des Skip-Gram-Modells, welche ebenfalls Subwortinformationen einbettet. So wurden, mit dem semantischen Fokus von Skip-Gram und den Subwortinformationen, gepaart mit der Schnelligkeit von FastText, zwei Modelle trainiert, eins für den AfD- und eins für den Grünen-Korpus. Die Wahl der Hyperparameter fand empirisch statt, hat aber in dieser Arbeit keinen Fokus, weshalb sich stattdessen auf Ergebnisse der verwandten Arbeiten bezogen wurde. So wurden also 80-dimensionale FastText-Modelle mit einer Window-Size von 7, um genügend Kontext einzubeziehen, einer Negative-Sample-Rate von 7 und einer Downsampling Rate von 1e-5, um besonders häufig auftretende Worte nicht überproportional oft zu betrachten, für 200 Epochen trainiert. Die Zahl der Epochen richtet sich nach den Ergebnissen von Komatsuzaki (2019), da nach tagelangem Experimentieren mit mehr Epochen keine Verbesserung der Ergebnisse stattfand und eine weitere, rigorosere Hyperparameteroptimierung den Rahmen dieses Reports sprengen würde. Wichtig ist hier ebenfalls zu nennen, dass eine „Verbesserung“ in diesem Fall nur subjektiv betrachtet werden kann, da es sich um unüberwachtes Lernen handelt. Besonders bei Embeddings sagt der Verlust auf den Trainingsdaten nämlich nicht sonderlich viel über die Güte der Lösung aus, vielmehr wird die Performanz auf gelabelten Daten betrachtet (siehe Mikolov et al. (1,3), 2013, Analogie-Tasks). Da für diesen speziellen Anwendungsfall allerdings keine gelabelten Daten bereitstanden, konnte die Performanz nur subjektiv anhand der ausgegebenen Wort-Nachbarschaften evaluiert werden. Als „Sanity-Check“ wurden dennoch zwei Analogie-Tasks ausgesucht, auf denen die eben genannten Modelle die besten Antworten lieferten. Mehr dazu in Kapitel 5. Alle Tokens, die aus weniger als 20 Zeichen bestehen, wurden beim Training ignoriert. Der effektive AfD-Korpus beträgt dann nur noch 6055 Tokens, der effektive Grünen-Korpus 5333.

4.4 Strukturelle Analyse

Um nun schließlich die Struktur der Embedding-Räume zu untersuchen, wurden mehrere Techniken angewandt. Zuerst wurden zu einigen ausgewählten

Schlüsselworten, die im relevanten Zeitraum zu den großen politischen Themen gehörten, die Nachbarschaften untersucht und miteinander verglichen: Wird das Schlüsselwort ungefähr im gleichen Kontext aufgefasst? Wo finden sich Unterschiede?

Anschließend wurden die Worte bestimmt, welche ausschließlich in nur einem Korpus vorhanden sind, um auf Diskrepanzen in der Wortnutzung und Thematisierung bestimmter Bereiche schließen zu können.

Schließlich wurde versucht, die größten strukturellen Diskrepanzen zwischen den Embeddings zu ermitteln. Dazu wurden die Embeddings aneinander ausgerichtet und die Worte ermittelt, deren Lage sich in den beiden Embedding-Räumen am meisten unterscheidet. Hierbei wurde sich größtenteils an der Methode von Azarbonyad et al. (2017) orientiert, die Embeddings wurden allerdings, statt Gradientenabstiegverfahrens, mittels des orthogonalen Procrustes aneinander ausgerichtet und die Stabilität eines Wortes durch die euklidische Distanz zwischen den Repräsentationen in den beiden Embedding-Räumen definiert. Um die Vorgehensweise von Azarbonyad et al. (2017) zu adaptieren mussten die Stoppworte allerdings wieder eingefügt werden, die FastText-Modelle wurden also auf leicht anderen Korpora trainiert. Nach dem Alignment wurden die Stoppworte wieder entfernt, in der Hoffnung die Qualität der Lösungen so verbessern zu können.

5. Ergebnisse

In diesem Kapitel werden die Ergebnisse mit der Vorgehensweise aus Kapitel 4 genannt. Für jeden Task wurden „Sanity Checks“ in Form von zwei Analogie-Tasks durchgeführt, ein Modell besteht den Test, wenn mindestens einer davon richtig gelöst wurde, die Lösung also in der Top-10-Nachbarschaft lag. Die Analogie-Tasks zur Verifizierung der Modellgüte lauteten:

Putin – Russland + Frankreich = ? (Lösung: Macron)

FDP – Wirtschaft + Umwelt = ? (Lösung: Grüne)

Das Modell zu Bündnis90/Die Grünen hat einen der Analogie-Tasks erfolgreich bestanden, das Modell zum AfD-Korpus beide. Höher-dimensionale Embeddings lieferten keine reproduzierbar besseren Ergebnisse.

5.1 Schlüsselworte und ihre Nachbarschaften

Die Bestimmung der Schlüsselworte relevanter politischer Themen der Jahre 2018 bis 2022 erfolgte mithilfe des Lebendigen Museums Online (LeMO, 2024), ein Kooperationsprojekt der Stiftung Deutsches Historisches Museum, der Stiftung Haus der Geschichte der Bundesrepublik Deutschland und des Bundesarchivs.

Die Nachbarschaften der ausgewählten Worte finden sich im Anhang in Tabellen 2 und 3 und beinhalten die zehn ähnlichsten Worte, gemessen an der Kosinus-Ähnlichkeit. Abb. 2 und 3 zeigen t-SNE Plots der Nachbarschaften, die besonders interessant erscheinen.

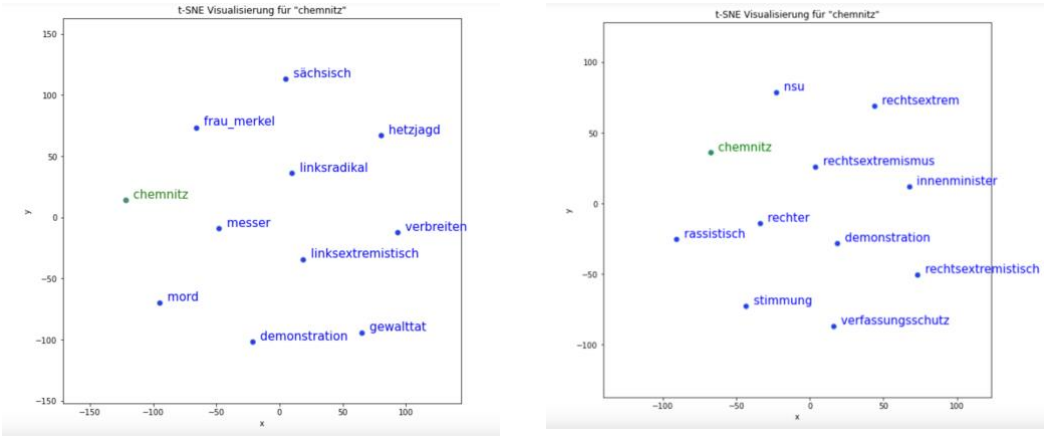

Abbildung 2: t-SNE Plots Chemnitz, Links: AfD, Rechts: Bündnis90/Die Grünen

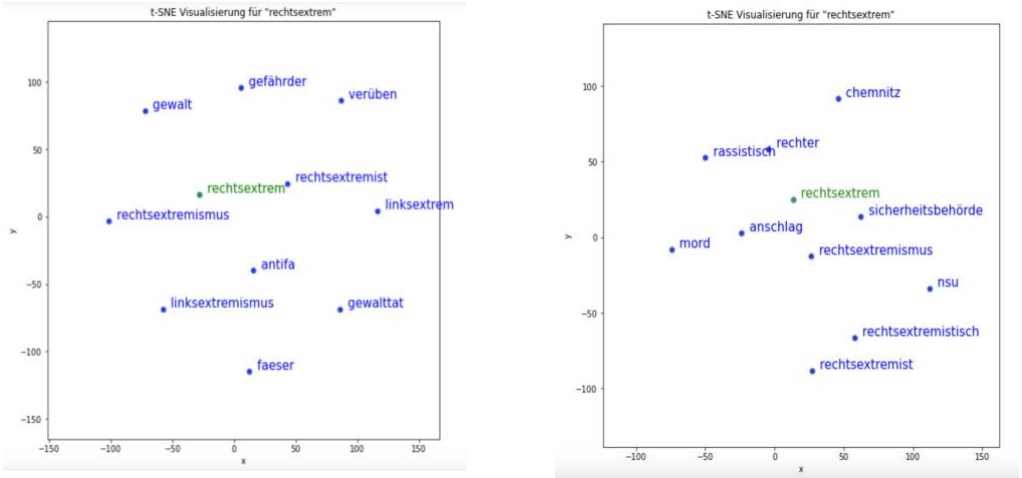

Abbildung 3: t-SNE Plots Rechtsextrem, Links: AfD, Rechts: Bündnis90/Die Grünen

Auf den ersten Blick ist ersichtlich, dass sowohl „Chemnitz“ als auch „Rechtsextrem“ in einen extremistischen Kontext eingeordnet wurden. Als Hintergrund sind hier die Ausschreitungen in Chemnitz im Jahr 2018 zu nennen, bei denen am Rande des Chemnitzer Stadtfestes ein Mann durch eine Messerattacke

getötet und zwei weitere schwer verletzt wurden. Rechtsextreme Gruppen riefen im Zuge dessen aufgrund von Nachrichten, welche den Flüchtlingsstatus und Migrationshintergrund der Täter publizierten, zu Demonstrationen auf. Es kam im weiteren Verlauf zu mehreren gewalttätigen Angriffen durch Rechte auf Personen mit vermeintlichem oder tatsächlichem Migrationshintergrund, Pressevertreter, Gegendemonstranten und ein jüdisches Restaurant (Brandau, Deutschlandfunk, 2018; Schulze, bpb, 2022). Als Reaktion fand im September 2018 ein Konzert vieler großer Künstler mit linkspolitischem Hintergrund unter dem Namen „Wir sind mehr“ statt, das entschieden als Zeichen gegen rechts galt (ZEIT ONLINE, 2018).

Vor diesem historischen Hintergrund lassen sich auf den zweiten Blick allerdings erhebliche Unterschiede in der kontextuellen Auffassung der Ausschreitungen in Chemnitz erkennen: Für die Grünen sind „Chemnitz“ und „Rechtsextrem“ sehr eng miteinander korreliert, sie sind gegenseitige Nachbarn. Auch in der Region um „Chemnitz“ finden wir fast ausschließlich Worte, welche mit Rechtsextremismus in Verbindung stehen („NSU“, „Rassistisch“, …). Die AfD hingegen scheint „Chemnitz“ und „Rechtsextrem“ nur am Rande oder gar nicht miteinander zu verbinden. Das einzige Wort in der Top-10-Nachbarschaft um „Chemnitz“, welches auf eine Verbindung hindeuten könnte, ist „Hetzjagd“. Dieses Wort findet sich mehrfach im Sprachgebrauch der AfD, oft auch in Verbindung mit Anschuldigungen an die restlichen Parteien, die Meinungsfreiheit einschränken zu wollen. Mithilfe einer Wortsuche nach „Chemnitz“ finden sich im AfD-Korpus allerdings zumeist nur sarkastische, oft auch verleugnende Statements:

> Sowohl die Kanzlerin als auch ihr Sprecher verbreiteten die Fake News, in Chemnitz sei es zu Hetzjagden auf Ausländer gekommen. (Alexander Gauland, Rede vor dem Bundestag, 12.09.2018)
>
> Die Bundesregierung schadet uns Deutschen im Ausland, weil sie Hetzjagden sieht, wo gar keine sind. (Steffen Kotré, Rede vor dem Bundestag, 27.09.2018)
>
> Es gab keine Hetzjagd. (Leif Erik Holm, Rede vor dem Bundestag, 27.09.2018)
>
> Gab es also keine Hetzjagd? Doch – nur nicht in Chemnitz, sondern auf Maaßen. (Gottfried Curio, Rede vor dem Bundestag, 27.09.2018)

Stattdessen finden sich in der Top-10-Nachbarschaft um „Chemnitz“ Worte wie „Messer“, „linksextremistisch“ und „linksradikal“, was davon zeugt, dass in den Reden eher weniger die rechtsextremen Ausschreitungen und mehr die linkspolitischen (hier als linksextrem betitelt) Reaktionen thematisiert wurden.

Auch die Tatsache, dass das Wort „Messer“ in so enger Verbindung mit „Chemnitz“ zu stehen scheint, gibt Hinweise darauf, dass die Messerattacke im Vorlauf der rechtsextremen Ausschreitungen viel mehr im Fokus der Reden von AfD-Politikern lag.

Weiterhin finden sich im Umfeld von „rechtsextrem“ bei der AfD auch Worte, die mit Linksextremismus beziehungsweise anderen radikalen Überzeugungen („Gefährder“) in Verbindung stehen, was daran liegen könnte, dass das FastText-Modell hier eher „Extremismus“ aufgefasst hat und „Rechtsextremismus“ im Vergleich mit dem Grünen-Modell weniger fein aufgefasst wird.

5.2 Exklusive Wortverwendungen und ihre semantische Aufgabe

Die exklusiv verwendeten Worte geben einen Einblick in die Themen, die der einen Partei wichtig, der anderen Partei eher unwichtig sind. Sie können sowohl Hinweise auf Themenfelder geben, die von besonderem Interesse für die Partei sind, als auch auf grundlegende Meinungsverschiedenheiten bezüglich der Wichtigkeit bestimmter Themen. In Tabelle 4 (siehe Anhang) sind die am meisten verwendeten exklusiven Worte pro Partei aufgelistet.

Zu erkennen ist hier, dass in Reden der AfD die häufig verwendeten exklusiven Worte oft in der Parteipolitik negativ konnotierte Themen beinhalten („asylbewerber“, „sozialismus“, „links_grün“). Es wird also oft ein überspitztes Feindbild thematisiert, das in Teilen der Politik der Regierungsfraktionen in etwa entspricht. Zusätzlich wird oft vom nationalen, bürgerlichen Interesse gesprochen („deutsch_steuerzahler“, „deutsch_volk“, „deutsch_interesse“, „bürgerlich“, „unser_bürger“), es wird sich also dem deutschen Volke angenommen, die Partei verschreibt sich scheinbar den Interessen des Bürgertums. Außerdem finden Hyperbeln („billion_euro“, „mikrogramm“, „rund_milliarde“) übermäßigen Gebrauch im Vergleich mit den Ergebnissen zur Grünen-Partei. Diese Beobachtungen entsprechen zusammengefasst laut Bundeszentrale für politische Bildung (kurz: bpb) der Definition des Populismus:

> Als Populismus bezeichnet man eine politische Grundhaltung, die in radikaler Opposition zu den herrschenden politischen und gesellschaftlichen Eliten steht und für sich selbst reklamiert, den „wahren“ Volkswillen zu erkennen und zu vertreten. Kern dieser Haltung ist die dichotomische Abgrenzung des moralisch guten, tugendhaften Volkes von den als korrupt und selbstsüchtig bezeichneten Vertretern des sogenannten Establishments. (Decker, F. bpb. Handwörterbuch politisches System - „Populismus“)

Interessant ist hierbei, dass das Wort „Populismus“ im AfD-Wahlprogramm zur Bundestagswahl 2021 kein einziges Mal Gebrauch findet, die Anzeichen populistischer Denkmuster finden sich dennoch auch darin wieder (AfD Wahlprogramm, 2021).

Im Sprachgebrauch der Grünen finden sich neben den ausschlaggebenden Parteithemen („klimakrise“, „ausbau_erneuerbar, „co2_preis“, „fossil_energie“) tendenziell eher Worte, die mit Fortschritt, progressivem Gedankengut und dem Aufruf zu aktivem Handeln assoziiert werden („aufbruch“, „verantwortung_übernehmen“, „dafür_einsetzen“, „gemeinsam_europäisch“, „anpacken“). Das entspricht in etwa den proklamierten Interessen der Partei:

> „Deutschland hat große Herausforderungen zu bewältigen: die ökologische Modernisierung der Wirtschaft, mehr soziale Gerechtigkeit und Anerkennung, mehr Zusammenhalt in der Gesellschaft und ein starkes Europa. Dafür muss nach Jahren einer Politik im Dauerkrisenmodus Weitsicht und Vorsorge einziehen. Nötig ist eine vorausschauende Politik, die Krisen verhindert und Mut macht, die nötigen Veränderungen anzugehen.“ (Wahlprogram Bündnis90/Die Grünen, 2021)

Im Vergleich ist auch die Nutzung von Hyperbeln und Superlativen niedriger. Dieser Unterschied zwischen einer klar proklamierten Parteiideologie und einer, die zwischen den Zeilen steht und in den Grundideen nur indirekt vermittelt wird, könnte Hinweise geben auf eine Rhetorik, die radikale Themen salonfähig zu machen versucht.

5.3 Größte strukturelle Diskrepanzen

Die automatische Bestimmung der größten strukturellen Diskrepanzen hat sehr viele Worte entdeckt, die bereinigten Top 5 sind in Tabelle 5 zu finden. Wichtig ist hier anzumerken, dass durch das nötige Inkludieren der Stoppwörter leicht andere Korpora als bei den anderen beiden Methoden verwendet wurden. Einzelne Worte wurden ausgefiltert, da die Nachbarschaften eine zu niedrige Qualität aufwiesen.

Diese Methode leidet generell unter eher ungenauen Darstellungen der Nachbarschaften, mit teilweise scheinbar unkorrelierten Worten in den Top 10 um ein Schlüsselwort. Dennoch lassen sich einige interessante Trends erkennen, die auch mit den vorherigen Beobachtungen übereinstimmen. So wird beispielsweise bei näherer Betrachtung der Nachbarschaft von „Familiennachzug“ in Abb. 4 klar, dass in Reden der Grünen-Partei eher humanitäre Ansichten vertreten werden („genfer“(_konvention), „geschützt“, „schutzsuchend“, „härtefall“), während in Reden der AfD das Thema eher bürokratisch und neutral bis hin zu negativ

konnotiert betrachtet wird („illegal“, „pass“, „abgelehnt_asylbewerber“, „migrant“). Diese Assoziationen weisen auf Vorurteile und unterschiedliche Grundeinstellungen gegenüber bestimmten Menschengruppen hin, was sich auch in der Politik der beiden Parteien eindeutig manifestiert.

Worte mit Bezug zur Abschiebung von Flüchtlingen finden sich in beiden Wortwolken, was daran liegen könnte, dass es seit der Flüchtlingskrise sehr viele Diskussionen zu den Szenarien gibt, in welchen eine Abschiebung gerechtfertigt wäre (siehe Abb. 5). Die kompletten Nachbarschaften der Top 5 unterschiedlichsten Worte lassen sich ebenfalls in Tabelle 5 einsehen.

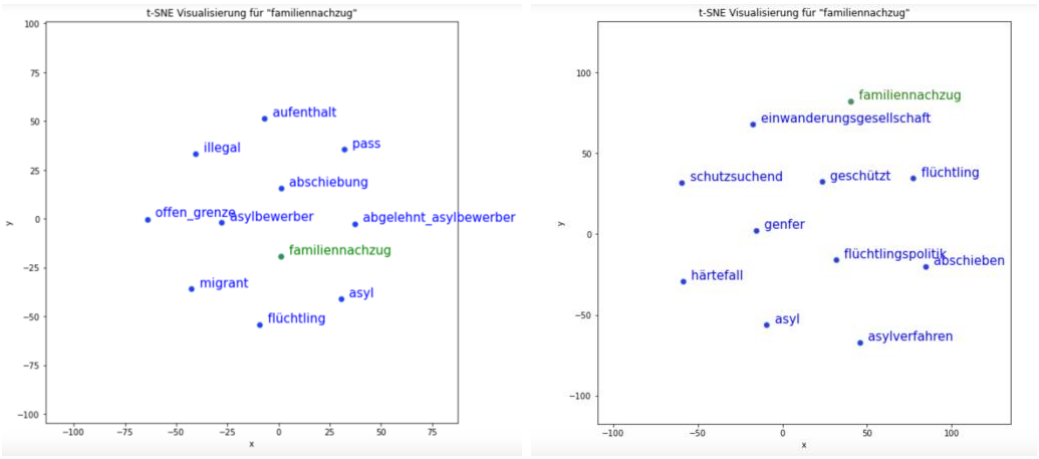

Abbildung 4: t-SNE Plots "Familiennachzug", Links: AfD, Rechts: Bündnis90/Die Grünen

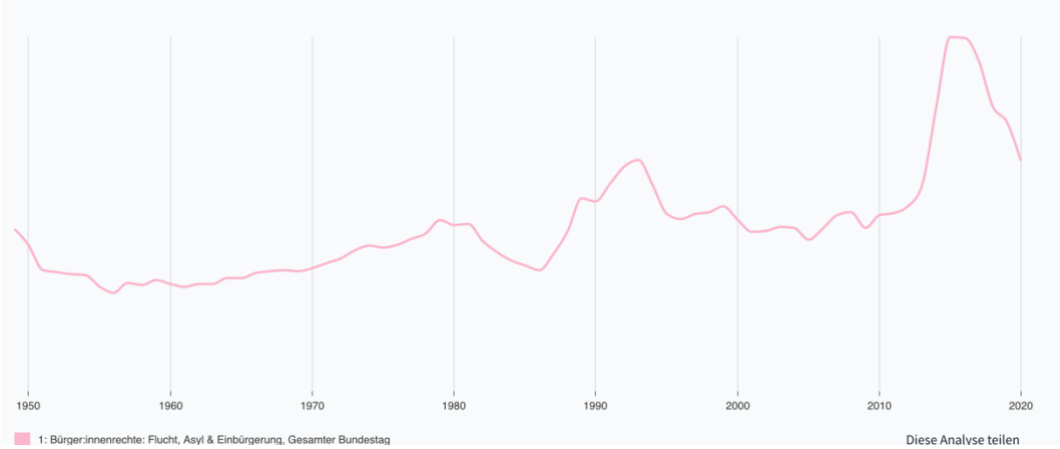

Abbildung 5: Diskursanalyse zum Thema "Bürger:innenrechte: Flucht, Asyl & Einbürgerung", Gesamter Bundestag[1]

6. Einschränkungen

6.1 Daten

Jede:r Data Scientist:in weiß: Die Qualität der Daten ist von größter Wichtigkeit. Für diese Arbeit standen nur Daten ohne Labels zur Verfügung, es gab keine Analogie-Validationen und keine geprüften Test-Datensätze für diese domänenspezifischen Anwendungen. Somit bestand Alles in Allem keine Möglichkeit, eine objektive Einschätzung der Güte der Embeddings und Verfahren zu treffen. Weiterhin lässt die effektive Korpusgröße für beide Parteien zu wünschen übrig, was zum Teil der Tatsache zu schulden ist, dass die AfD erst seit Ende 2017 Mitglied des Bundestages ist und deshalb nur über einen relativ kurzen Zeitraum Reden zur Verfügung standen. Wie Mikolov et al. (2013, (3)) allerdings schon feststellten, sind die Word2Vec-Modelle sehr datenhungrig, die Performanz lässt sich also vermutlich mit mehr Daten verbessern.

6.2 Transferlernen

Ein weiteres Problem, mit den eben genannten verwandt, ist die nicht garantierte Ähnlichkeit der Embedding-Räume, welche für das Alignment von großer Bedeutung ist. Das liegt vermutlich an der Spezifizität der Korpora, sie können die komplette deutsche Sprache einfach nicht vollständig auffassen, da sie domänenspezifisch nur im politischen Kontext entstandene Texte beinhalten. Eine Möglichkeit, das zu lösen, besteht darin, ein vortrainiertes FastText-Modell weitergehend auf den Expertendaten zu trainieren. Dieses Vorgehen wird *Finetuning* genannt. Die Wortvektoren, die bereits bestehen und geprüft sind, würden dann durch die injizierten Kontexte in eine bestimmte Richtung verschoben werden, sich also der politischen Überzeugung des Korpus anpassen. Die FastText API (Bojanowski et al., 2017), die für dieses Projekt genutzt wurde, unterstützt allerdings nur die überwachte Version des Finetunings und ein eigenes FastText-Modell für die gesamte deutsche Sprache zu trainieren hätte den Rahmen dieser Arbeit und meine Ressourcen bei Weitem gesprengt.

[1] https://opendiscourse.de/l/Ihj7S

7. Fazit und Aussicht

Die kontextuelle Auffassung von bestimmten Streitthemen trägt in einer Gesellschaft, in der das Verbreiten von Falschinformationen zunehmend leichter wird, einen sehr hohen Stellenwert. Um diese kontextuelle Auffassung zu bestimmen können Embeddings herangezogen werden. In diesem Report wurde festgestellt, dass in der Tat einige allgemeinbekannten politischen Interessen und damit auch die kontextuellen Auffassungen dieser nachvollziehbar repräsentiert und untersucht werden können. So wurden Hinweise darauf entdeckt, dass die politischen Überzeugungen der Alternative für Deutschland empirisch nachvollziehbar dem Populismus zugeordnet werden können, auch wenn das nur indirekt aus dem Wahlprogramm hervorging. Des Weiteren wurden durch die semantische Bedeutung bestimmter Schlüsselworte am Beispiel der Ausschreitungen in Chemnitz die parteiliche Rezeption ebendieser festgestellt und im Korpus bestätigt. Das Vergleichen der Struktur von Embeddings bietet somit einen vielversprechenden Anhaltspunkt für diese Analysemethoden, wenn denn auch genügend Daten vorliegen. Denn das automatisierte Verfahren benötigt wesentlich mehr Daten um den Ergebnissen eine akzeptable Aussagekraft verleihen zu können. Hier wurden eher wenig aussagekräftige Wortwolken erzeugt, wodurch eine qualitative Analyse nur bedingt möglich war. Doch auch hier könnten eine größere Datenmenge und das Anwenden von Transferlernen zusammen mit den restlichen Punkten aus Kapitel 6 Besserung verschaffen.

8. Anhang

Tabelle 2: Top-10-Nachbarschaften Bündnis90/Die Grünen, dim=80, epoch=200, loss=1.46

Thema	Top-10-Nachbarschaft	Thema	Top-10-Nachbarschaft
Fahrverbot	hardwarenachrüstung, verkehrsminister, blau, herr_scheuer, fahrzeug, diesel, grenzwert, nachrüstung, pkw, sauber	Covid	pandemie, virus, infektion, erkrankung, pandemiebekämpfung, intensivstation, folge, krankheit, impfstoff, eindämmung
Trump	usa, us, donald_trump, multilateralismus, amerikanisch, politik, welt, eigen, multilateral, europäisch_union	Impfpflicht	impfen, impfung, allgemein, impf, abwägung, impfquot, herbst, variante, infektion, impfkampagne
Asyl	schutzsuchend, flucht, geflüchtet, geflüchteter, flüchtlingspolitik, genfer(_konvention), familiennachzug, unwürdig, abschottung, flüchtling	Rechtsextrem	rechtsextremismus, rechtsextremistisch, rechter, mord, sicherheitsbehörde, chemnitz, anschlag, rechtsextremist, nsu, rassistisch
Türkei	türkisch, erdogan, völkerrechtswidrig, syrien, einmarsch, rüstungsexport(e), außenminister, völkerrecht, maas	Feminismus	feministisch, feministisch_außenpolitik, reproduktiv, entwicklungszusammenarbeit, entwicklungspolitik, selbstbestimmt, geschlechtergerechtigkeit, selbstbestimmen, außenpolitik, gleichberechtigung
Chemnitz	rechtsextrem, stimmung, rechtsextremismus, rechter, remonstration, innenminister, rechtsextremistisch, nsu, rassistisch, verfassungsschutz	Europa	europäisch, europäisch_union, gemeinsam_europäisch, europäer, gemeinsam, deutschland, macron, kontinent, solidarität, frieden

Tabelle 3: Top-10-Nachbarschaften AfD, dim=80, epoch=200, loss=1.29

Thema	Top-10-Nachbarschaft	Thema	Top-10-Nachbarschaft
Fahrverbot	fahrverbote, grenzwert, stickstoffdioxid, stuttgart, mikrogramm, umwelthilfe, auto, diesel, fahrzeug, scheuer	Covid	corona, virus, impfstoff, infektion, pandemie, lockdowns, krankheit, erkrankung, märz, pcr
Trump	donald, vereinigt_staat, amerikanisch, usa, us, amerikaner, außenminister, iran, washington, russland	Impfpflicht	impfen_lassen, impfung, impfen, geimpft, grundrecht, impf, herr_lauterbach, ungeimpft, impfzwang, impfstoff
Asyl	flüchtling, grenze, migrant, einreisen, migration, familiennachzug, asylbewerber, integration, einwanderung, illegal	Rechtsextrem	rechtsextremismus, rechtsextremist, linksextrem, faeser, verüben, antifa, linksextremismus, gewalt, gefährder, gewalttat
Türkei	erdogan, türkisch, syrien, kurde, türke, konflikt, beitritt, bergkarabach, armenien, diplomatisch	Feminismus	feministisch, gender, verschleiern, anträg, worüber, aktionismus, wild, kümmern, außenpolitik, strukturell
Chemnitz	hetzjagd, sächsisch, demonstration, mord, verbreiten, messer, linksextremistisch, linksradikal, gewalttat, frau_merkel	Europa	europäisch, nation, eu, frieden, frankreich, großbritannien, souveränität, deutsch, kontinent, europäer

Tabelle 4: Meistverwendete exklusive Worte, ab mind. 90 Erwähnungen, in Klammern die Anzahl der Erwähnungen im gesamten Korpus

Partei	Auszug der meistverwendeten exklusiven Worte
AfD	afd_fraktion (557), linker (381), sozialistisch (257), kernenergie (219), deutsch_steuerzahler (184), alternative_deutschland (181), links_grün (178), kernkraftwerk (177), deutsch_volk (172), deutsch_interesse (166), sozialismus (155), antifa (149), sozialsystem (147), rot_grün (136), sed (136), altpartei (133), asylbewerber (124), planwirtschaft (122), billion_euro (113), fremd (112), bürgerlich (111), entwicklungshilfe (107), unser_bürger (102), zwang (96), heimisch (95), mikrogramm (95), genosse (92), rund_milliarde (92), co2_steuer (91)
Bündnis90/Die Grünen	klimakrise (617), ausbau_erneuerbar (203), studierend (194), geflüchtet (165), aufbruch (129), groß_herausforderung (126), wichtig_schritt (108), co2_preis (108), bündnis_grüne (106), fossil_energie (104), verantwortung_übernehmen (100),

	kindergrundsicherung (99), demokratisch_fraktion (95), brauchen_dringend (94), dafür_einsetzen (93), kinderzuschlag (93), patientinn_patient (92), darüber_diskutieren (92), gemeinsam_europäisch (91), anpacken (90)

Tabelle 5: Worte mit den größten strukturellen Diskrepanzen, folgende Worte wurden aufgrund von mangelnder Nachbarschaftsqualität aussortiert (dürr, tendenz, anleihe, verzögerung, frauenhäuser)

Wort	Distanz in Embeddings (auf 4 Nachkommastellen gerundet)	Nachbarschaften (Afd-Modell)	Nachbarschaften (Grünen-Modell)
wiederaufbau	1.6064	syrien, syrisch, assad, zivil, bürgerkrieg, militärisch, irak, abziehen, bundeswehr, heer	krisenbewältigung, bahnhof, jüdin, solidarität, berühren, signal, rehberg, historisch, vergessen, europa
Italien	1.5822	spanien, frankreich, italienisch, griechenland, eu, italiener, ezb, europäisch, euro, tschechien	spanien, griechenland, hafen, türkei, de, großbritannien, europa, flüchtling, frankreich, mitgliedstaat
familiennachzug	1.5717	migrant, flüchtling, offen_grenze, pro_monat, abgelehnt_asylbewerber, gefährder, asylbewerber, illegal, einwanderer, verschlimmern	härtefall, genfer, drittstaat, geduldet, asylverfahren, einschränkung, syrien, abschottung, abschieben, högl
ausspielen	1.5704	gegeneinander, autofahrer, verkehrsträger, verkehrsmittel, verkehrspolitik, belassen, tierschutz, strecke, verkehrswende, schiene	gegeneinander, stellung, umweltschutz, naturschutz, spielen, klimaschutz, deal, spaltung, überhaupt_nichts, rücken
bewältigung	1.5638	bewältigen, effizient, ungleich, eindämmung, krise, sicherung, aussetzung, substanz, nachhaltigkeitsziel, wirtschaftlich	krise, bewältigen, meistern, pandemie, herausforderung, begegnen, modernisierung, groß_herausforderung, entschlossen, handeln

Literaturverzeichnis

Alternative für Deutschland. (2021). Wahlprogramm zur Bundestagswahl 2021. https://www.afd.de/wp-content/uploads/2021/06/20210611_AfD_Programm_2021.pdf (zuletzt aufgerufen am 21.08.2024 um 13:20 Uhr)

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, & Jaap Kamps. (2017). Words are Malleable: Computing Semantic Shifts in Political and Media Discourse.

Russa Biswas, Mehwish Alam, & Harald Sack. (2020). Is Aligning Embedding Spaces a Challenging Task? A Study on Heterogeneous Embedding Alignment Methods.

Piotr Bojanowski, Edouard Grave, Armand Joulin, & Tomas Mikolov. (2017). Enriching Word Vectors with Subword Information.

Bundesregierung. (22.05.2024). Woran sie Falschmeldungen erkennen können. https://www.bundesregierung.de/breg-de/schwerpunkte/umgang-mit-desinformation/falschmeldungen-erkennen-1750146 (zuletzt aufgerufen am 19.8.2024, 18:50 Uhr)

Brandau, B. (31.12.2018). Chemnitz – eine zerrissene Stadt. Deutschlandfunk. https://www.deutschlandfunk.de/2018-in-sachsen-chemnitz-eine-zerrissene-stadt-100.html (zuletzt aufgerufen am 21.08.2024 um 17:41)

Bündnis90/Die Grünen. (2021). Wahlprogramm zu Bundestagswahl 2021. https://www.gruene.de/artikel/wahlprogramm-zur-bundestagswahl-2021 (zuletzt aufgerufen am 21.08.2024 um 14:15 Uhr)

Decker, F. bpb. Handwörterbuch politisches System „Populismus“ (zuletzt aufgerufen am 21.08.2024 um 17:58 Uhr)

Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Gavagai AB. A Brief History of Word Embeddings | Gavagai. Gavagai. https://www.gavagai.io/text-analytics/a-brief-history-of-word-embeddings/, zuletzt aufgerufen am 17.08.2024

William L. Hamilton, Jure Leskovec, & Dan Jurafsky. (2018). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.

Haus der Geschichte. Lebendiges Museum Online. https://www.dhm.de/lemo/ (zuletzt aufgerufen am 20.08.2024 um 15:35 Uhr)

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Herve Jegou, & Edouard Grave. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion.

Alexander Kalinowski, & Yuan An. (2020). A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs.

Aran Komatsuzaki. (2019). One Epoch Is All You Need.

Limebit GmbH. (2021). OpenDiscourse. https://opendiscourse.de/ (zuletzt aufgerufen am 21.8.2024 um 17:17)

MDR. (11.12.2023). Prozess zu Ausschreitungen in Chemnitz 2018: Zwei Angeklagte fehlen (zuletzt aufgerufen am 21.08.2024 um 11:50 Uhr)

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, & Jeffrey Dean. (2013). Distributed Representations of Words and Phrases and their Compositionality.

Tomas Mikolov, Kai Chen, Greg Corrado, & Jeffrey Dean. (2013). Efficient Estimation of Word Representations in Vector Space.

Tomas Mikolov, Quoc V. Le, & Ilya Sutskever. (2013). Exploiting Similarities among Languages for Machine Translation.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Association for Computational Linguistics.

Schulze, C. (2022). Rechtsextremismus. Bundeszentrale für politische Bildung (zuletzt aufgerufen am 21.08.2024 um 11:30 Uhr)

ZEIT ONLINE. (4.09.2018). Alles, was Anstand hat (zuletzt aufgerufen am 20.08.2024 um 13:45 Uhr)

Selbstständigkeitserklärung

Ich versichere hiermit, dass die vorliegende Hausarbeit mit dem Thema

_____

_____

von mir selbst und ohne jede unerlaubte Hilfe angefertigt wurde, dass sie weder an einer anderen Hochschule noch an dieser Universität als Prüfung vorgelegen hat und dass sie weder ganz noch in Auszügen veröffentlicht worden ist. Die Stellen der Prüfungsleistung – einschließlich Tabellen, Karten, Abbildungen usw. –, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

_____

Vor- und Nachname in Druckbuchstaben

_____

Ort, Datum, Unterschrift